Electronic Version
Stylesheet Version v1.1.1

Description

Knowledge Discovery Apparatus and Method

Federal Research Statement

[0001] This invention was developed entirely under the internal efforts of EagleForce Associates, Inc. ("EagleForce"). EagleForce claims all rights with regard to this patent.

Background of Invention

[0002] 1. Field of the Invention

[0003] This invention is directed to an apparatus and method for performing knowledge discovery by extracting elements of information that are useable to an analyst with regard to an area of inquiry, whether or not that inquiry has been formally framed or the "inquiry" is generated by the apparatus in the course of automated processes.

[0004] 2. Description of the Related Art

[0005] There are many applications performing Knowledge Discovery (KD), ranging from Federal and Defense intelli-

gence to business intelligence.

[0006] Often, in such applications, many KD tools are used to perform specific steps in the KD process as identified in Claims (1) through(7). More recently, various suites of such tools have been assembled to perform sequences of related KD operations. An example of such is the architecture adopted for the (2002) Joint Intelligence Virtual Architecture system. These systems are limited by the lack of either a Feedback Loop or a Utility Function modifying the Feedback Loop.

SUMMARY OF INVENTION

[0007] This invention overcomes the above-noted disadvantages. An apparatus in accordance with this invention is constructed to receive data feeds from one or more data sources, where the data feeds may include live and / or stored data, including "structured" (database) data, unstructured (e.g., document, web page), semi-structured (e.g., military Commander's Intent orders, militaryFrag(mentation) orders, or military or commercial email), along with audio, video, and / or image data. It is the intent of described metataggng methodology and apparatus to provide the highest and best use of the indexing, classification, and categorization of information resi-

dent within the collateral networks. The distinguishing feature of the methodology is the use of the "EF Feedback Loop", a process that incorporates the highest and best use of multiple COTS tools. The feedback loop is a widely accepted calibration concept, commonly deployed in this environment for elements of ranking algorithms, type weights, and type proximity-weights. The feedback loop is used in conjunction with one or more of the EF Utility Function(s). The purpose of the utility functions is to iteratively adjust the parameter controls sent back via the feedback loop process in order to maximize results according to a given benefit or utility.

[0008] The primary challenges associated with retrospective metadata tagging are:

[0009] 1.Creating the right metadata "concept classes" that identify those corpus elements (e.g. documents, pages, paragraphs) containing inquiry-relevant concepts, and

[0010] 2.Ensuring scalability.

[0011] The issue of scalability compels us to use an architectural suite of integrated COTS tools as integral to the apparatus, along with the control mechanisms of feedback loops governed by utility functions. This is the only means by which metadata tagging can be retrospectively done, while

still maintaining the ability to handle very large (e.g., order-of-terabyte, or $O(10^{12})$, sized corpora.

[0012] The scalability issue also motivates us to use an integrated COTS suite to reduce the manpower overhead and minimize the level of human interaction required to support the retrospective markup process, while still maintaining the quality of the metadata markup needed for precision searching.

[0013] The key issue in controlling scalability, and in reducing manpower overhead, is to determine correct parameter settings governing the metadata tagging process as well as information retrieval in response to metatag-based queries. This is undoubtedly the most significant challenge in the data analysis and metatagging process. One reason that this is so challenging is that when retrospective metadata tagging is introduced as an additional processing stage on top of preliminary data metatagging, the issues associated with corpus size and scalability are exacerbated. Thus, it is crucial to find a method by which metadata tagging can be done, both initially and retrospectively, in a manner that both makes precise inquiry possible and which allows scaling to very large corpora.

[0014] Google patent holders, Drs. Sergey Brin and Lawrence

Page, who in their paper "The Anatomy of a Large-Scale Hypertextual Web Search Engine," state, "Figuring out the right values for these parameters is something of a black art", express the importance of this challenge.

[0015] Like most others, Drs. Brin and Page place the user as the initial and primary element(s) of the feedback loop. There, the "user may optionally evaluate all of the results that are returned." But it is precisely this positioning that becomes untenable as very large corpora are considered. This "Google" process, common among most COTS tagging and search products, has clearly achieved less than satisfactory results in the challenging intelligence data-parsing environment. Even user-oriented search training functions ultimately only serve to constrain results based on the limitations of a particular tool's mathematical capabilities.

[0016] To enhance this well-established query process into structured, unstructured, and semi- structured data, many in the Defense, Intelligence, and commercial environment have begun developing suites of tools that utilize different algorithms against the same data set. Two major issues evolve when using such suites:

[0017] 1. Query results using these suites generally differ based on the order of the data flow.

[0018] 2. The results are extremely inconsistent and become virtually unusable as the data corpora expand.

[0019] The latter issue of results inconsistency is directly related to the issue of scalability, which is a primary concern when dealing with retrospective metadata tagging. Generally, the metaschema between the tools is unique to the individual product and integration, even that which extends to the API level, allowing the individual tool to read and optimize its portion of the metadata. Knowledge is organized and presented in an extremely robust manner when the data corpora are small. However, as the size of the originating file expands, the discovery of relevant knowledge and entities/concepts to tag, suffers greatly.

[0020] The EagleForce approach to minimizing the user interaction level required for precise searching is to first define a functional architecture in which different levels of knowledge representation and knowledge processing are used in successive manner. Both *initial* and *retrospective* metadata tagging are done at Level 1. Higher levels allow for different degrees of correlation among the data. When these correlations are done, it is possible to generate focused and pertinent retrospective metadata tagging directives. This is done partially through modifying the ranking func-

tion that guides metadata tagging. The modified ranking function is used to present the rank impact of the change on all previous searches.

[0021] Here the EF FeedBack Loop runs a Level 1 classifier tool at a very simple level as a first pass. This serves to focus on getting those documents that have the highest, richest data relative to the inquiry as we position our classifier to operate with a very tight sigma - i.e., a document has to have lots of hits on very simple, core keywords in order to be selected and moved forward. For this purpose, we use a Bayesian classifier with Shannon relevance ranking. The value of the EF Feedback Loop and the EF Utility function allows the use of multiple independent or collective Level 1 tools. The EF Feedback Loop and the EF Utility Function apparatus is employed to control the processing limits without affecting fidelity by disbursing the workflow to multiple reasoning parsers.

[0022] Once the initial Level 1 pass is complete, the EF Feedback Loop and Utility Function allow the user to set the number and/or relevance scale to the first order of Level 2. The system will automatically push the most relevant sources to Level 2 so as to allow that portion of the system to apply its independent "noun phrase" parsing and "co-

occurrence" algorithms to the classification/ categorization process. The Level 2 processor will then push only its new classification/categorization concepts back to level 1 for re-indexing. Following the second pass the EF Feed-Back Loop and its associated Utility functions allows the second pass to Level 2 to take its most relevant data to Level 3 for its independent "verb" parsing algorithms. New concepts or classifications are passed back from Level 2 and to Level 1 for re-indexing and with results returned to Level 2. The EF Feedback Loop has now allowed 5 sets of algorithms to apply 3 independent sets of metadata markings that are all read in their entirety, in exactly the same fashion by the integrated system prior to the user seeing the first query result.

[0023] The EF Feedback Loop is controlled by a set of "Utility Functions" which are designed to support the centralization of information technology services that are of common concern to the Intelligence Community. This methodology employs the indexing schema in the same manner for structured and unstructured data, however we employ the specific use of structured data OLAP tools to address the EF FeedBack Loop independently from the noun phrase or verb parsing.

Brief Description of Drawings

[0024] FIG 1 Illustrates the challenge of scalability, which shows how very large data corpora must be processed in order for to extract meaning relative to a given inquiry.

[0025] FIG. 2 is exemplary schematic views of the seven levels for a complete KD architecture includes five representation levels (1 through 5) and two control levels (6 and 7), in accordance with the invented method and apparatus. This figure shows the EagleForce "Representation Levels" concept, which is a foundation for building a knowledge discovery architecture. Levels 1 through 5 are detailed with Level 0 indexing (not shown) being reserved for the ingestion of extremely large data sets. Level 6 provides feedback control of lower levels, and Level 7 contains a utility function that is used to optimize feedback. This scalability serves to significantly enrich the metatagging process.

[0026] FIG. 3 is provides a schematic view of data flow through the apparatus, including the optional step 0, but not reflecting optional step 5c, beginning with the original data corpus and the transformation of the data corpus through the operations performed upon the data corpus.

Detailed Description

[0027] DESCRIPTION OF THE ARCHITECTURE

[0028] The method and apparatus consists of a tiered set of representation levels, herein described as five representation levels, along with an optional Level 0, together with the EF FeedBack Loop methodology and the EF Utility Function, which is designed to index, classify, and categorize data at eight levels of processing. The preferred embodiment is to employ a COTS-based architecture, making use of "best of the breed"

[0029] existing and proven tools. This embodiment has, in cooperation with several COTS vendors, developed and already demonstrated an integrated architecture with essential capabilities from Levels 1 through4. he addition of the technology provided by a Level 5 capability will complete the basic suite. Note that within this architectural framework, there is typically more than one COTS capability. Within the overall architectural concept, it is possible to use a customer-preference for a specific COTS product within a given appropriate level, or to use more than one COTS capability, again within a given level.

[0030] The EF FeedBack Loop begins with the order of scalability assuming that the incoming data set is on the order of 1 terabyte. The first order of business is to determine the

time interval (Day, Month) to provide a consistent measurement basis for evaluation. The approach allows the first order of indexing (identification of documents with key words) to be metatagged as they are found in the document without the generalization into classes, concepts, co-occurrence-, etc. This level is used as the heavy lift, which allows the system and not the user to initiate the definition process as to whether a document has any potential relevance whatsoever, or if it can just be tossed. The goal at Level 0 is to reduce the amount of data as much as possible, without losing anything potentially useful.

[0031] The preferred embodiment for this method and apparatus is based on a "Plug and Play" mindset. Thus, both the method and the apparatus are agnostic with respect to database vendor. A similar approach is employed throughout the architecture for the apparatus.

[0032] INTERFACE DESCRIPTION WITHIN THE ARCHITECTURE

[0033] There are two different classes of interfaces within the architecture. The first, and generally more straightforward, is the passing of data and metadata between tools. This apparatus and method solves the associated interface problems between several different tools, usually by a

combination of special interface code at the API level, and use of intelligence in tool-specific metadata. Additional tools can be integrated as necessary.

[0034] The second interface type involves passing of control between applications. This method and architecture has solved this via the EF Feedback Loop and the EF Utility Functions. The EF Feedback Loop has been described in the said claim (6). The EF Utility Functions are a set of measures of the value (utility) of an intermediate or final output to the end-user, and have been described in the said claim (7). Utility functions thus provide a metric by which a proposed feedback action can be measured, and the overall performance of the system improved. Multiple utility functions are typically required because there are several independent axes that may be used to determine effectiveness.

[0035] ADVANTAGES AND BENEFITS OF THE METHOD AND APPARATUS

[0036] This method and apparatus provide multiple benefits to the end user. Since the architecture comprehends the value of common look and feel, the usual difficulties in switching from tool to tool are mitigated. As capability is added, an increasing number of queries can be formed in

natural language (English). In addition to facilitating ease of use and productivity, both of these factors reduce the amount of training required to employ these capabilities. Addition of a vector-based geo-referencing capability will enable the user to "drill down"based on geospatial locality.

[0037] Advantageously, the invented apparatus and method can be used to preferentially extract relatively sparse concept classes and most especially various combinations of concept classes (where each "concept class" can be expressed as a category, a set of nouns and / or noun phrases, or a single noun or noun phrase, depending on the embodiment of the invention) along with identification of the relationships (single or multiple verbs, or verb sets) linking different concept classes. At the same time, the influence of "contextual" information can be incorporated to preferentially refine a given concept class, or to add more information relative to an area of inquiry. As an example, including geo-spatial references at Level 4 of the processing allows for "neighborhoods" surrounding a given occurrence to be preferentially tagged via feedback into the Level 1 process. Similarly, use of a Language Variant method at Level 4 can be used to identify geospatial re-

gions of interest when a name of interest (found during Level 1 or Level 2 processing) is identified and then one or more Language Variants of that name are identified in Level 4. If occurrences of these proper name Language Variants are then found as a result of feedback into a lower level (e.g., Level 1), then the geospatially-referenced regions associated with the Language Variants provide context for later iterations of the feed forward process that begins at Level 1.

[0038] These together with other features and advantages, which will become subsequently apparent, reside in the details of construction and operation of the invented apparatus and method as more fully hereinafter described and claimed, reference being made to the accompanying drawing, forming a part hereof, wherein like numerals refer to like parts throughout the view.